RECOGNIZING THE NUMERIC LANGUAGE IN NATURAL SPOKEN DIALOGUE

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates to a system for numeric language recognition in natural spoken dialogue.

2. Description of the Related Art

Speech recognition is a process by which an unknown speech utterance (usually in the form of a digital PCM signal) is identified. Generally, speech recognition is performed by comparing the features of an unknown utterance to the features of known words or word strings. Hidden Markov models (HMMs) for automatic speech recognition (ASR) rely on high dimensional feature vectors to summarize the short-time, acoustic properties of speech. Though front-ends vary from speech recognizer to speech recognizer, the spectral information in each frame of speech is typically codified in a feature vector with thirty or more dimensions. In most systems, these vectors are conditionally modeled by mixtures of Gaussian probability density functions (PDFs).

Recognizing connected digits in a natural spoken dialog plays a vital role in many applications of speech recognition over the telephone. Digits are the basis for credit card and account number validation, phone dialing, menu navigation, etc.

Progress in connected digit recognition has been remarkable over the past decade. For databases recorded under carefully monitored laboratory conditions, speech recognizers have been able to achieve less than 0.3% word error rate. Dealing with telephone speech has added a new dimension to this problem. Variations in the spectral characteristics due to different channel conditions, speaker populations, background noise and transducer

equipment cause a significant degradation in recognition performance. Previous practice has strictly focused on dealing with constrained input speech to produce digit sequences.

SUMMARY OF THE INVENTION

In accordance with the principles of the invention, the set of words or phrases that are relevant to the task of understanding and interpreting number strings is referred to as the "numeric language". The "numeric language" defines the set of words or phrases that play a key role in the understanding and automation of users' requests. According to an exemplary embodiment of the invention, the numeric language consists of the set of word or phrase classes that are relevant to the task of understanding and interpreting number strings, such as credit card numbers, telephone numbers, zip codes, etc., and consists of six distinct phrase classes including "digits", "natural numbers", "alphabets", "restarts", "city/country name", and "miscellaneous".

In the exemplary embodiment of the invention, a system includes a speech recognition processor that receives unconstrained fluent input speech and produces a string of words that can include a numeric language, and a numeric understanding processor that converts the string of words into a sequence of digits based on a set of rules. An acoustic model database utilized by the speech recognition processor includes a first set of hidden Markov models that characterize the acoustic features of numeric words, a second set of hidden Markov models that characterize the acoustic features of the remaining vocabulary words, and a filler model that characterizes the acoustic features of out-of-vocabulary utterances. An utterance verification processor verifies the accuracy of the string of words. A validation database stores a grammar, and a string validation processor outputs validity information based on a comparison of the sequence of digits with the grammar. A dialogue manager processor initiates an action based on the validity information.

Other aspects and advantages of the invention will become apparent from the following detailed description and accompanying drawing, illustrating by way of example the features of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates a numeric language recognition system in accordance with the principles of the invention; and

FIG. 2 illustrates an acoustic model database in accordance with the principles of the invention.

DETAILED DESCRIPTION

For a better understanding of the invention, together with other and further objects, advantages, and capabilities thereof, reference is made to the following disclosure and the figures of the drawing. For clarity of explanation, the illustrative embodiments of the present invention are presented as comprising individual functional blocks. The functions these blocks represent may be provided through the use of either shared or dedicated hardware, including, but not limited to, hardware capable of executing software. For example, the functions of the blocks presented in FIG. 1 may be provided by a single shared processor. Illustrative embodiments may comprise digital signal processor (DSP) hardware, read-only memory (ROM) for storing software performing the operations discussed below, and random-access memory (RAM) for storing DSP results. Very large scale integration (VLSI) hardware embodiments, as well as custom VLSI circuitry in combination with a general purpose DSP circuit, may also be provided. Use of DSPs is advantageous since the signals processed represent real physical signals, processes and activities, such as speech signals, room background noise, etc.

This invention is directed to advancing and improving numeric language recognition in the telecommunications environment, particularly the task of recognizing numeric

words when embedded in natural spoken dialog. In particular, the invention is directed toward the task of recognizing and understanding users' responses when prompted to respond with information needed by an application involving the numeric language, such as, for example, their credit card or telephone number. We have identified those words that are relevant to the task and enhance the performance of the system to recognize those relevant words.

By way of example, and not limitation, in a specific embodiment of the invention, the numeric language forms the basis for recognizing and understanding a credit card and a telephone number in fluent and unconstrained spoken input. Our previous experiments have shown that considering the problem of recognizing digits in a spoken dialogue as a large-vocabulary continuous speech recognition task, as opposed to the conventional detection methods, can lead to improved system performance.

In an exemplary system for recognizing the numeric language in a natural spoken dialogue, illustrated in FIG. 1, a feature extraction processor 12 receives input speech. A speech recognition processor 14 is coupled to the feature extraction processor 12. A language model database 16 is coupled to the speech recognition processor 14. An acoustic model database 18 is coupled to the speech recognition processor 14.

A numeric understanding processor 20 is coupled to the speech recognition processor 14. An utterance verification processor 22 is coupled to the speech recognition processor 14. The utterance verification processor 22 is coupled to the numeric understanding processor 20. The utterance verification processor 22 is coupled to the acoustic model database 18.

A string validation processor 26 is coupled to the numeric understanding processor 20. A database 28 for use by the string validation processor 26 is coupled to the string validation processor 26.

A dialog manager processor 30 is coupled to the string validation processor 26. The dialogue manager processor 30 initiates action according to the invention in response to the results of the string validation performed by the string validation processor 26.

Using a spoken dialogue system imposes a new set of challenges in recognizing digits, particularly when dealing with naive users of the technology. In this example, during a spoken dialogue users are prompted with various open questions such as, "What number would you like to call?", "May I have your card number please?", etc. The difficulty in automatically recognizing responses to such open questions is not only to deal with fluent and unconstrained speech, but also to be able to accurately recognize an entire string of numerics (i.e., digits or words identifying digits) and/or alphabets. In addition the system ought to demonstrate robustness towards out-of-vocabulary words, hesitation, false-starts and various other acoustic and language variabilities.

Performance of the system was examined in a number of field trial studies with customers responding to the open-ended prompt "How may I help you?" with the goal to provide an automated operator service. The purpose of this service is to recognize and understand customers' requests whether it relates to billing, credit, call automation, etc.

In an important part of the field trials, customers were prompted to say a credit card number or a telephone number to obtain call automation or billing credit. Various types of prompts were studied with the objective to stimulate maximally consistent and informative responses from large populations of naive users. These prompts are engineered towards asking users to say or repeat their credit card or telephone number without imposing rigid format constraints.

The system is optimized to recognize and understand words in the dialogue that are salient to the task. Salient phrases are essential for interpreting fluent speech. They are commonly identified by exploiting the mapping from unconstrained input to machine action.

Those salient phrases that are relevant to the task are referred to as "numerics." Numeric words and phrases in the numeric language are the set of words that play a key role in the understanding and automation of customers' requests. In this example, the numeric language consists of six distinct phrase classes including digits, natural numbers, alphabets, re-starts, city/country name, and miscellaneous.

Digits, natural numbers and alphabets are the basic building blocks of telephone and credit card numbers. Users may say "my card number is one three hundred fifty five A four…". Restarts include the set of phrases that are indicative of false-starts, corrections and hesitation. For example, "my telephone number is nine zero eight I'm sorry nine seven eight…". City/country names can be essential in reconstructing a telephone number when area codes are missing. For example, "I would like to call Italy and the number is three five…". Finally, there are a number of miscellaneous phrases that can alter the sequencing of the numbers. Such phrases are "area-code", "extension number", "expiration date", etc. For our application, the numeric language consisted of a total of one hundred phrases.

According to the invention, numeric recognition in spoken dialogue systems is treated as a large vocabulary continuous speech recognition task where numerics are treated as a small subset of the active vocabulary in the lexicon. The main components of the numeric recognition system illustrated in FIG. 1 are described as follows.

In the feature extraction processor 10, the input signal, sampled at eight kHz, is first pre-emphasized and grouped into frames of thirty msec durations at every interval of ten msec. Each frame is Hamming windowed, Fourier transformed and then passed through a set of twenty-two triangular band-pass filters. Twelve mel cepstral coefficients are computed by applying the inverse discrete cosine transform on the log magnitude spectrum. To reduce channel variation while still maintaining real-time performance, each cepstral vector is normalized using cepstral mean subtraction with an operating look-ahead delay of thirty speech frames. To capture temporal information in the signal, each normalized cepstral vector along

with its frame log energy are augmented with their first and second order time derivatives. The energy coefficient, normalized at the operating look-ahead delay, is also applied for end-pointing the speech signal.

Accurate numeric recognition in fluent and unconstrained speech clearly demands detailed acoustic modeling of the numeric language (the numeric words and phrases). It is essential to accurately model out-of-vocabulary words (the non-numerics) as they constitute over eleven percent of the database. Accordingly, our design strategy for the acoustic model 18 has been to use two sets of subword units. Referring to FIG. 2, a first set 36 of hidden Markov models (HMMs) that characterize the acoustic features of numeric words is dedicated for the numeric language. A second set 38 of HMMs that characterize the acoustic features of the remaining vocabulary words is dedicated for the remaining vocabulary words. Each set 36, 38 applies left-to-right continuous density hidden Markov models (HMMs) with no skip states.

In the first set 36 dedicated for recognition of numerics, context-dependent acoustic units have been used which captured all possible inter-numeric coarticulation. The basic structure is that each word is modeled by three segments; a head, a body and a tail. A word generally has one body, which has relatively stable acoustic characteristics, and multiple heads and tails depending on the preceding and following context. Thus, junctures between numerics are explicitly modeled. Since this results in a huge number of subword units, and due to the limited amount of training data, the head-body-tail design was strictly applied for the eleven digits (i.e., “one”, “two”, “three”, “four”, “five”, “six”, “seven”, “eight”, “nine”, “zero”, and “oh”). This generated two hundred seventy-four units which were assigned a three-four-three state topology corresponding to the head-body-tail units, respectively.

The second set 38 of units includes forty tri-state context-independent subwords that are used for modeling the non-numeric words, which are the remaining words in the vocabulary. Therefore, in contrast to traditional methods for digit recognition, out-of-vocabulary

words are explicitly modeled by a dedicated set of subword units, rather than being treated as filler phrases.

To model transitional events between numerics, non-numerics and background/silence, an additional set 40 of units is used. Three filler models with different state topologies are also used to accommodate for extraneous speech and background noise events. In total, three hundred thirty-three units are employed in the exemplary units. Each state includes thirty-two Gaussian components with the exception of the background/silence model which includes sixty-four Gaussian components. A unit duration model, approximated by a gamma distribution, is also used to increment the log likelihood scores.

The language model database 16 is used by the speech recognition processor 14 to improve recognition performance. The language model database 16 contains data that describes the structure and sequence of words and phrases in a particular language. In this specific example, the data stored in the language model database 16 might indicate that a number is likely to follow the phrase “area code” or that the word “code” is likely to follow the word “area”; or, more generally, the data can indicate that in the English language, adjectives precede nouns, or in the French language, adjectives follow nouns. While language modeling is known, the combination of the language model database 16 with the other components of the system illustrated in FIG. 1 is not known.

Speech, or language, understanding is an essential component in the design of spoken dialogue systems. The numeric understanding processor 20 provides a link between the speech recognition processor 14 and the dialogue manager processor 30 and is responsible for converting the recognition output into a meaningful query.

The numeric understanding processor 20 translates the output of the recognizer 14 into a “valid” string of digits. However, in the event of an ambiguous request or poor recognition performance, the numeric understanding processor 20 can provide several

hypotheses to the dialogue manager processor 30 for repair, disambiguation, or perhaps clarification.

A rule-based strategy for numeric understanding is implemented in the numeric understanding processor 20 to translate recognition results (e.g., N-best hypotheses) into a simplified finite state machine of digits only. Several classes of these rules which aim to translate input text into a digit sequence are presented in TABLE 1.

Rule	Definition	Example
Naturals	translating natural numbers	one eight hundred and two → 1 8 0 0 2
Restarts	correcting input text	nine zero eight sorry nine one eight → 9 1 8
Alphabets	translating characters	A Y one two three → 2 9 1 2 3
City/Country	translating city/country area codes	calling London, England → 4 4 1 8 8
Numeric Phrases	realigning digits	nine on two area code nine zero one → 9 0 1 9 1 2
Out-of vocabulary	filtering	what is the code for Florham Park → 9 7 3

TABLE 1

The utterance verification processor 22 identifies out-of-vocabulary utterances and utterances that are poorly recognized. The utterance verification processor 22 provides the dialogue manager 30 with a verification measure of confidence that may be used for call confirmation, repair or disambiguation. The output of the utterance verification processor 22 can be used by the numeric understanding processor 20.

Information is validated before being sent to the dialogue manager processor 30. Due to ambiguous speech inputs and possible errors in the dialogue flow, sometimes

customers' responses to prompts represent invalid telephone number or credit card numbers. Sometimes, even with a robust system, misrecognition occurs.

In order to alleviate this problem, and to improve system performance generally, task-specific knowledge is introduced. The task-specific knowledge can be in the form of grammars that correspond to national and international telephone numbers and/or various credit card numbers, for example.

In the exemplary system illustrated in FIG. 1, a set of valid credit card numbers and a set of valid telephone numbers are stored in the validation database 28 for use by the string validation processor 26. The string validation processor checks the validation database 28 to determine whether the sequence of digits output by the numeric understanding processor 20 corresponds to an existing telephone number or credit card number.

In the specific example illustrated in FIG. 1, the string validation processor 26 outputs validity information that indicates the validity of the sequence of digits produced by the numeric understanding processor 20. The validity information indicates a valid, partially valid, or invalid sequence of numbers.

Checking whether the sequence of digits at the output of the numeric understanding processor 20 corresponds to an existing telephone or credit card number is valuable information in two respects. First, it provides a type of rejection which may be used to narrow down the error rate. Second, it guarantees that a valid credit card or telephone number is being processed.

The validity information and the sequence of digits output from the numeric understanding processor 20 are passed to the dialogue manager processor 30. The dialogue manager processor 30 initiates one or more actions based on the sequence of digits and the validity information.

A characterization of the problem of recognizing digits embedded in a spoken dialog has been presented herein. The invention is useful in recognizing credit card numbers,

telephone numbers, zip codes, dates, times, etc. It will be appreciated that the principles of the invention are also applicable to pattern recognition generally.

While several particular forms of the invention have been illustrated and described, it will also be apparent that various modifications can be made without departing from the spirit and scope of the invention.